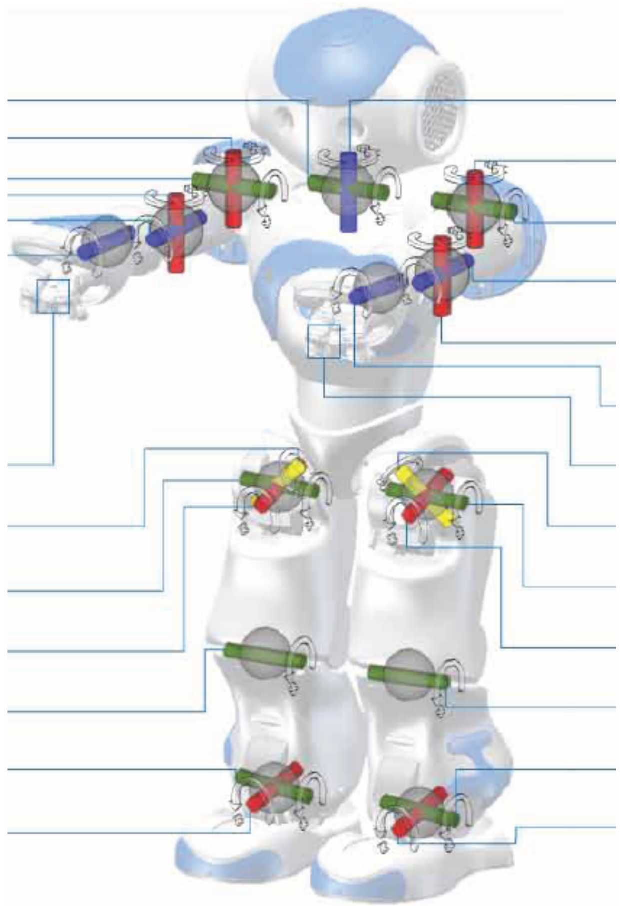
Decision Making
in Robots and Autonomous Agents

Causality:
How should a robot reason about cause and effect?

Subramanian Ramamoorthy
School of Informatics

13 February, 2018

What do you Need to Know about your Robot?



What does Robot Need to Know?

- Given access to raw data channels for various (uninterpreted) sensors and motors

- Devise a procedure for learning that will tell you what you need for various tasks (as yet unspecified)
  - What types of models?
  - What types of learning methods?

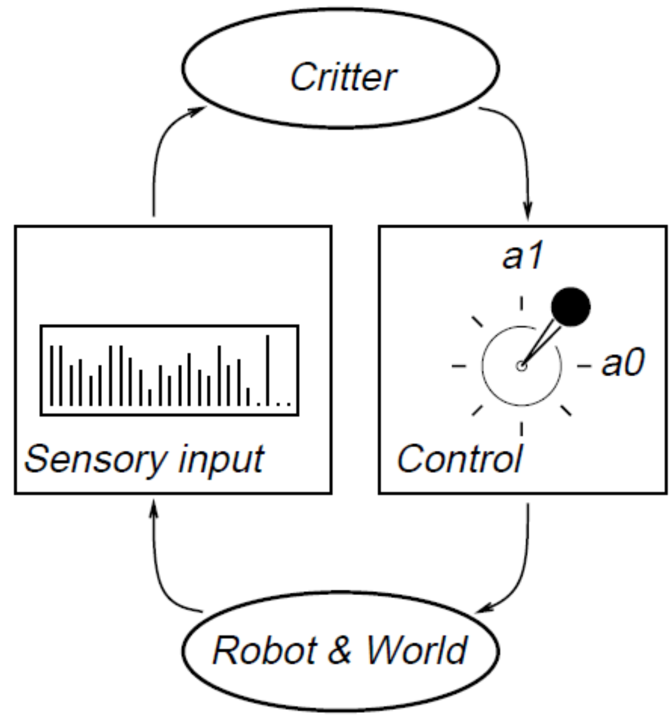
What are you Learning from?

1.896007	4.323841	22.664253	4.202899	1.213200
1.635462	1.698352	22.664253	2.294771	2.627882
1.894624	4.323150	22.664253	4.202899	1.213891
1.635462	1.698352	22.664253	2.293389	2.627882
1.894624	4.323150	22.664253	4.202899	1.211126
1.635462	1.699043	22.664253	2.290625	2.627882
1.895315	4.323150	22.664253	4.202899	1.211126
1.635462	1.699043	22.664253	2.289934	2.630646
1.895315	4.323150	22.664253	4.206354	1.210435
1.636153	1.699043	22.664253	2.294080	2.630646
1.895315	4.323150	22.664253	4.200825	1.211126
1.630624	1.699043	22.664253	2.318960	2.629264
4.375673	2.443358	22.664253	22.664253	1.119210
1.638917	1.725996	2.295462	22.664253	1.953368
22.664253	22.664253	1.250519	22.664253	22.664253
22.664253	2.322415	22.664253	22.664253	2.000363
22.664253	22.664253	1.148236	4.363925	1.913975
22.664253	2.323106	2.730856	2.043211	22.664253
22.664253	1.196613	4.400553	2.316196	2.352824
2.359735	22.664253	2.068782	22.664253	22.664253
1.259504	22.664253	1.965117	2.151714	22.664253
3.287191	22.664253	22.664253	22.664253	22.664253
22.664253	1.991378	22.664253	2.374248	22.664253
2.467546	22.664253	22.664253	22.664253	1.321011
22.664253	1.997598	2.197326	22.664253	2.168300
2.450960	22.664253	2.661054	22.664253	1.303734
22.664253	1.997598	2.194562	22.664253	2.168300
2.450960	22.664253	2.660363	22.664253	1.303734
22.664253	1.996216	22.664253	22.664253	2.171064
2.457871	22.664253	2.656908	22.664253	1.306498
1.244990	1.976174	2.004509	2.356970	22.664253
2.756426	22.664253	4.558124	22.664253	22.664253
1.193158	22.664253	1.908446	2.112321	2.002436
3.316217	2.086059	4.537391	3.118563	22.664253
22.664253	1.164823	1.886331	1.891860	2.328635
2.411567	2.788908	2.090206	22.664253	22.664253
22.664253	1.101242	4.308637	1.858687	2.087441
2.382541	2.786835	2.070855	4.485559	22.664253
22.664253	22.664253	1.065304	1.849012	2.227735
2.406730	22.664253	2.050813	4.466898	22.664253
22.664253	4.712930	1.023147	3.835233	2.220824
22.664253	22.664253	2.786143	2.019023	22.664253
22.664253	22.664253	1.087420	22.664253	3.797223
22.664253	2.428845	22.664253	1.991378	4.363234
22.664253	22.664253	22.664253	0.993430	1.832425
22.664253	2.428845	22.664253	22.664253	2.000363
22.664253	22.664253	4.680449	0.993430	22.664253
1.815839	22.664253	22.664253	22.664253	2.001054
22.664253	22.664253	4.681140	0.993430	22.664253
1.815839	22.664253	22.664253	22.664253	2.001054
22.664253	22.664253	4.691506	0.990666	22.664253
1.815839	22.664253	22.664253	22.664253	2.001054
22.664253	22.664253	4.680449	0.993430	22.664253
1.815839	22.664253	22.664253	22.664253	2.001054
22.664253	22.664253	4.676993	0.991357	22.664253
1.815839	22.664253	22.664253	22.664253	2.001054
22.664253	22.664253	4.682522	0.995503	22.664253
1.809619	22.664253	2.470311	22.664253	2.001745
4.424050	22.664253	22.664253	1.110226	22.664253
1.807546	22.664253	22.664253	22.664253	22.664253
2.054268	4.432344	22.664253	22.664253	22.664253
2.142729	1.689368	22.664253	2.464782	22.664253

22.664253	1.967881	2.245012	22.664253	2.397054
22.664253	1.968572	2.245012	22.664253	2.397745
22.664253	1.968572	2.245012	22.664253	2.397745
22.664253	1.967881	2.245012	22.664253	2.397745
22.664253	1.968572	2.245012	22.664253	2.397745
22.664253	1.965808	2.243630	22.664253	2.397745
3.963778	1.983085	2.162080	2.341766	22.664253
1.931944	2.409494	2.098499	22.664253	2.130981
2.146185	2.048740	22.664253	2.050813	1.548383
2.014876	22.664253	2.012802	1.486184	1.602980
22.664253	1.909829	1.419839	1.538708	22.664253
22.664253	1.552530	1.510373	22.664253	22.664253
1.945075	1.551148	1.501388	22.664253	2.250541
1.943001	1.551148	1.501388	22.664253	2.250541
2.026624	1.555294	1.508300	22.664253	2.252614
22.664253	1.930562	1.439881	1.555294	22.664253
2.182813	2.088132	22.664253	1.617493	22.664253
2.003818	22.664253	2.003818	1.526268	22.664253
2.016258	2.238101	2.102645	1.585703	1.697661
2.347986	2.269892	22.664253	1.788195	1.653431
2.101263	2.078457	2.312740	2.190415	1.671399
1.838645	2.135818	2.362499	22.664253	1.740509
1.832425	2.386688	2.105410	22.664253	2.242248
1.815839	3.524238	2.095043	2.344531	2.220133
1.815839	3.524238	2.095043	2.344531	2.217368
1.815839	3.524238	2.095734	2.345222	2.221515
1.815839	3.524238	2.095734	2.344531	2.224279
1.815839	3.524238	2.095734	2.345222	2.220824
1.814457	2.156551	2.094352	2.339693	2.217368
3.797223	1.822059	2.347986	2.070164	22.664253
1.003797	1.808237	1.823441	2.343148	2.301682

An Experiment: How Much can we Learn from Uninterpreted Data?

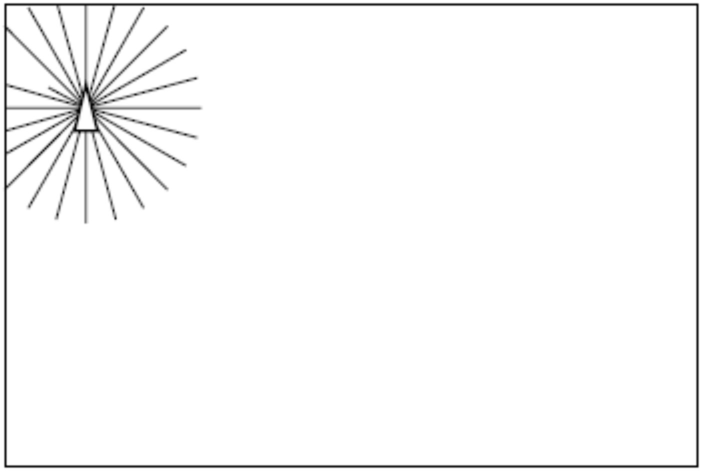
- Learn models of robot and environment with no initial knowledge of what sensors and actuators are doing
- Many learning methods begin this way, e.g., RL, but the goal here is to construct a representation incrementally and continually as well



[D. Pierce, B.J. Kuipers, Map learning with un-interpreted sensors and effectors, *Artificial Intelligence* 91:169-227, 1997.]

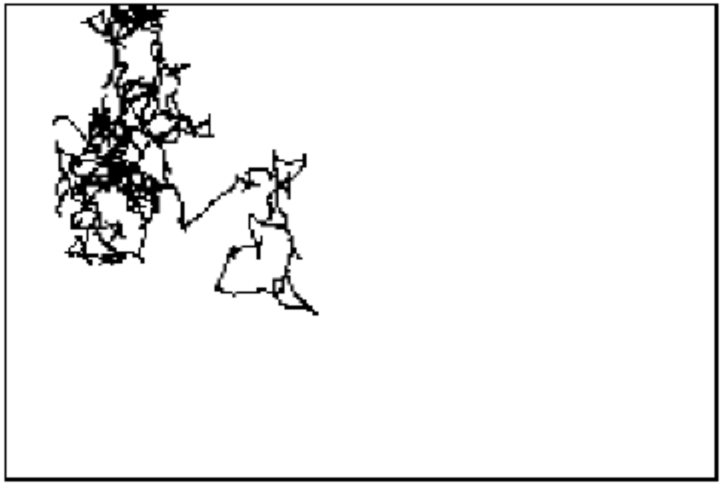
Simple Scenario

- Robot critter has a set of distance sensors (range) – one of which is defective – but it doesn't know that yet
- Other sensors: battery power, digital compass
- It has a track-style motor apparatus – turn by differentially actuating its wheels



What do you Learn from?

Randomized actions (hold a randomly chosen action for 10 time steps), repeatedly applied



How does environment appear in the data?
Can there be a simple empirical learning scheme?

One Step: Go from Raw Channels to Structure of Sensor Array

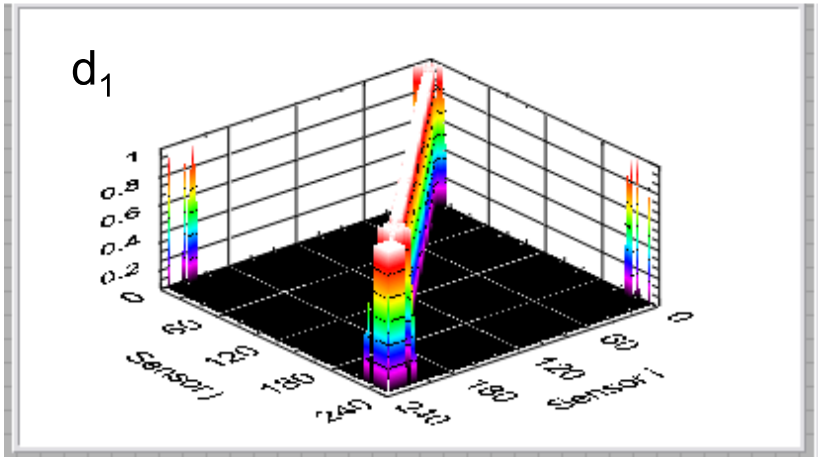
- Sensors may come in groupings: ring of distance sensors, array of photoreceptors, video camera, etc.
- We first want to extract groupings based on two criteria:
  - Sensors that have similar values over time
  - Sensors that have a similar frequency domain behaviour
- Two simple hypothesised distance metrics:

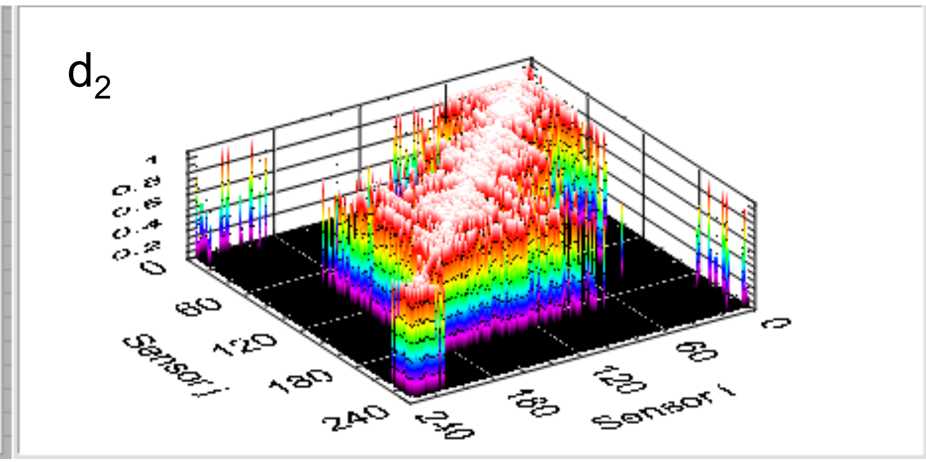
$$d_{1,ij}(t) = \frac{1}{t+1} \sum_{\tau=0}^{t} |x_i(\tau) - x_j(\tau)|.$$

$$d_{2,ij} = \frac{1}{2} \sum_{l} |(\boldsymbol{dist}\, x_i)_l - (\boldsymbol{dist}\, x_j)_l|,$$

Distribution (e.g., counts)

Example Trace





$$i \approx_k j \quad \text{if } d_{k,ij} < \min\{\varepsilon_{k,i}, \varepsilon_{k,j}\}.$$

$$\varepsilon_{k,i} = 2 \min_j \{d_{k,ij}\}.$$

Extending the Group Notion

We can reason transitively about similarity:

$$i \sim j \quad \text{iff} \quad i \approx j \vee \exists k\colon\ (i \sim k) \wedge (k \sim j).$$

So, a wandering trace might yield something like this as groups:

(0 1 2 22 23) (0 1 2 3 23) (0 1 2 3 4) (1 2 3 4 5) (2 3 4 5 6) (3 4 5 6 7) (4 5 6 7) (5 6 7 8 9) (7 8 9 10) (7 8 9 10 11) (8 9 10 11 12) (9 10 11 12 13) (10 11 12 13 14) (11 12 13 14 15) (12 13 14 15 16) (13 14 15 16 17) (14 15 16 17 18) (15 16 17 18 19) (16 17 18 19) (17 18 19 21) (20) (19 21 22 23) (0 21 22 23) (0 1 21 22 23) (24) (25) (26) (27) (28).

Upon Taking the *Transitive Closure*

(0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 21 22 23)

(20) *defective*

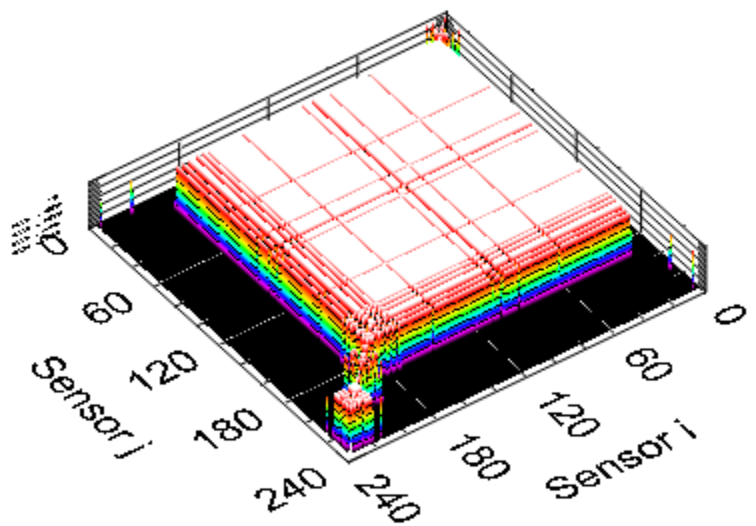
(24) *battery voltage*

(25) *east*

(26) *north*

(27) *west*

(28) *south*



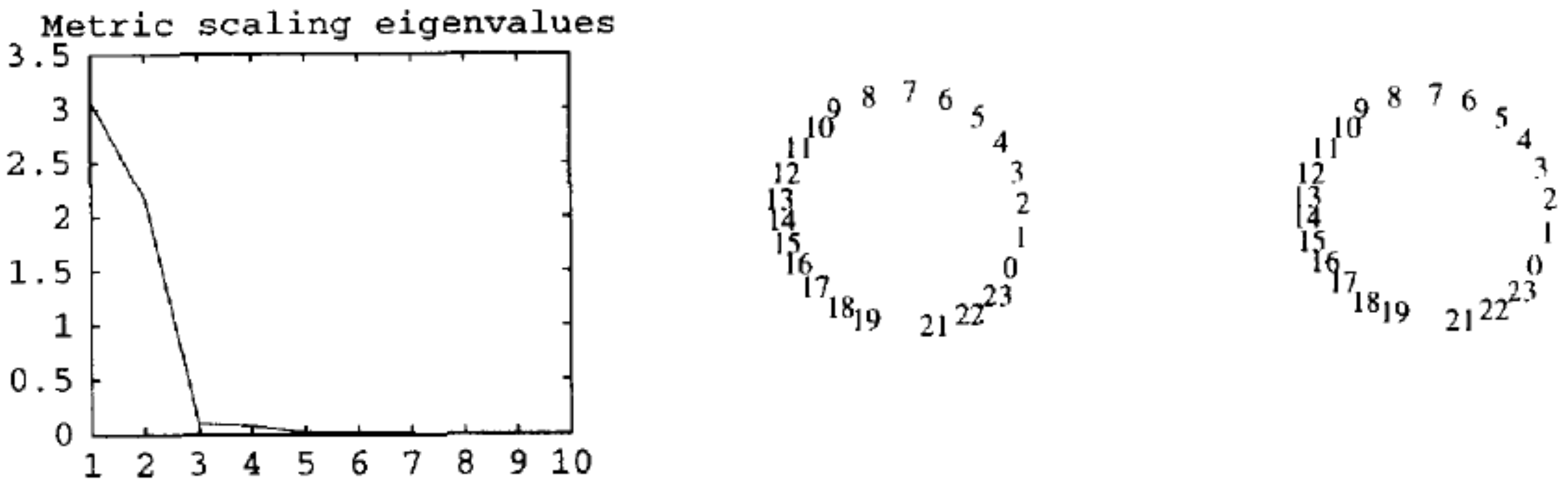
Getting at the Structure of Array

- Task is to find an assignment of positions (in space) to elements that captures the structure of the array as reflected in distance metric d_1.
- Distance between positions in image ≈ distance between elements according to d_1.

$$\|(\mathbf{pos}\ y_i) - (\mathbf{pos}\ y_j)\| = d_{1,ij},$$

- This is a constraint satisfaction problem: n sensor elements yield n(n-1)/2 constraints.
- Could solve by metric scaling: $E = \frac{1}{2}\sum_{ij}(\|(\mathbf{pos}\ y_i) - (\mathbf{pos}\ y_j)\| - d_{ij})^2.$

Structural Model of Distance Array

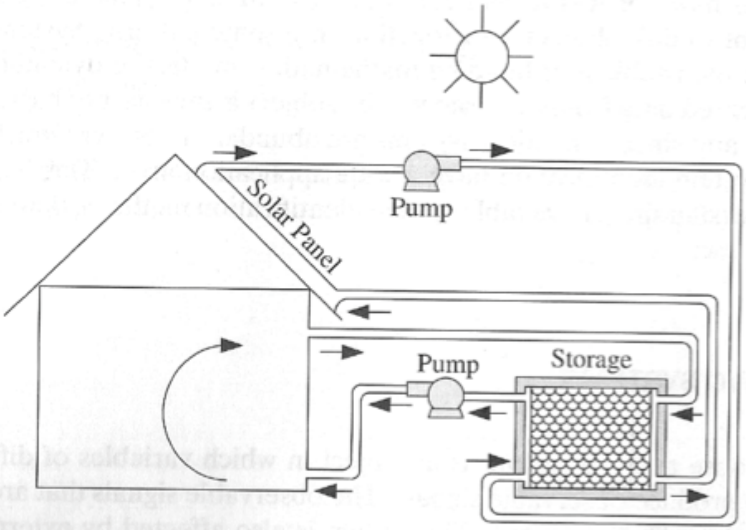


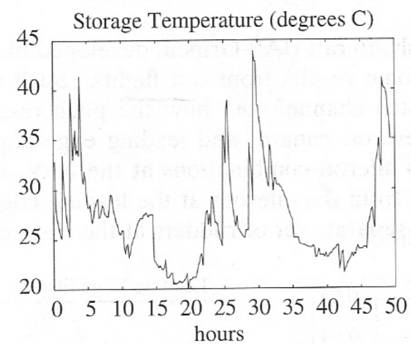
a b c

Various Types of Models

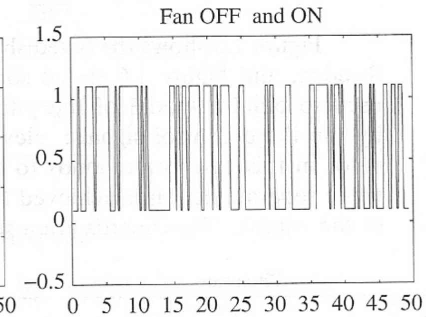
- Models of motion
  - Own dynamics
  - Object dynamics
  - Other agents
- Models of environment
  - Space & how I move in space
  - Other navigation considerations
- Models of self
  - What is the connection between my sensors and actuators?
  - What do the sensorimotor channels even mean?
  - How to ground all of the above at this low level?

Example: Solar-Heated House (Ljung)

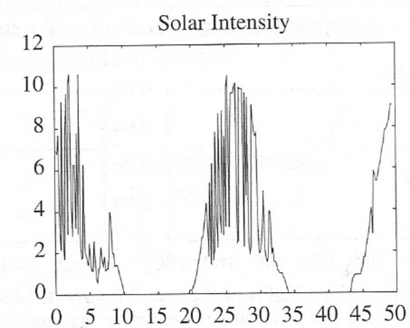




(a) Storage temperature



(a) Pump velocity



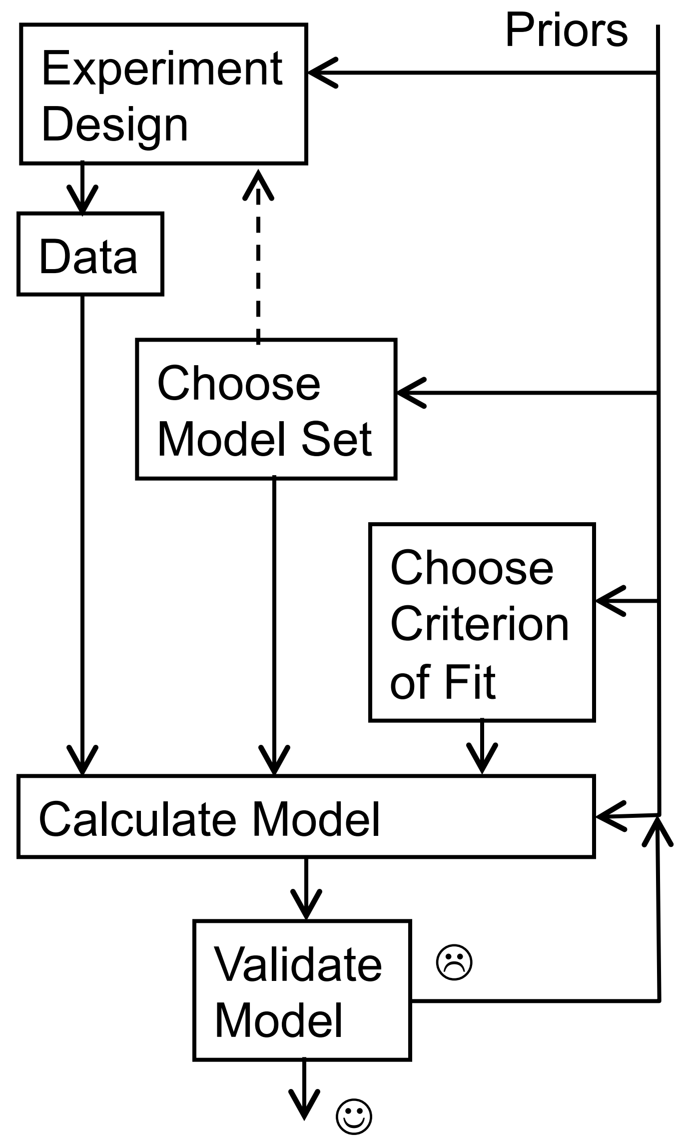
- The sun heats the air in the solar panels
- The air is pumped into a heat storage (box filled with pebbles)
- The stored energy can be later transferred to the house
- For control, one cares about how solar radiation, $w(t)$, and pump velocity, $u(t)$, affect heat storage temperature, $y(t)$.

System Identification in Engineering

In building a model, the designer has control over three parts of the process

1. Generating the **data set**
2. Selecting a (set of) **model structure** (e.g., autoregressive linear model)
3. Selecting the **criteria** (e.g., least squares over output error), used to specify the optimal parameter estimates

A very popular approach involves (recursive) parameter estimation



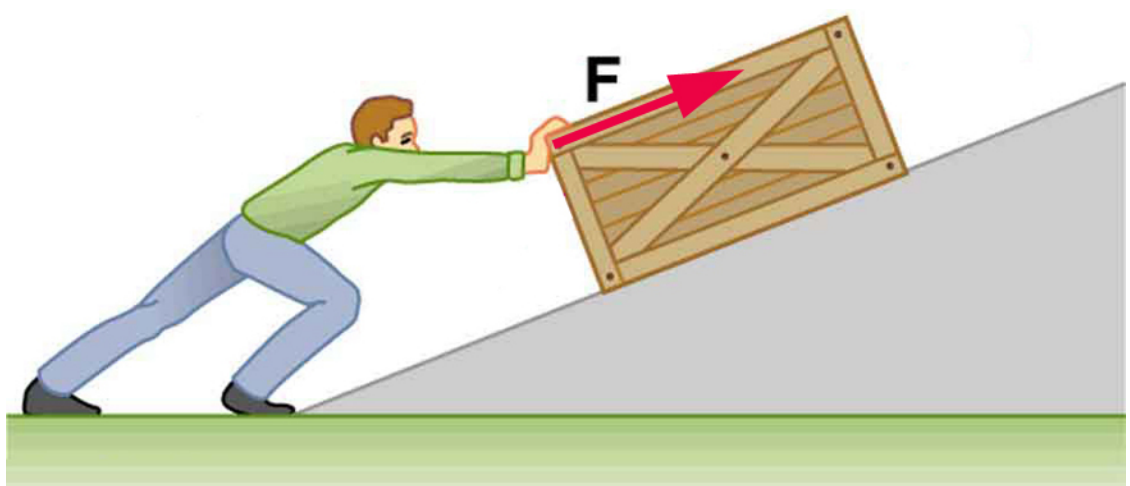
On the Nature of Scientific Questions

Science seeks to understand and explain physical observations

- Why doesn't the wheel turn?
- What if I make the beam half as thick, will it carry the load?
- How do I shape the beam so it will carry the load?

What Do Laws Tell Us About Causality?

- Does acceleration cause the force?
- Does the force cause the acceleration?
- Does the force cause the mass?



Different Views on Causation

- Hume (1711 – 1776) [Causation as perception]

We remember seeing the flame and feeling a sensation called heat; without further ceremony, we call one cause and the other effect

- Pearson (1857 – 1936) [Statistical Machine Learning view]

Forget causation! Correlation is all you should ask for.

- Pearl (1936 -) [Mathematics of causality]

Forget empirical observations! Define causality based on a network on known, physical causal relationships

Two Major Questions about Causality

1. Learning of causal connections: What empirical evidence legitimizes a cause – effect connection?
   – How do people ever acquire knowledge of causation
   – e.g., does a rooster cause the sun to rise?
   – succession, correlations are not sufficient
   – e.g. Roosters crow before dawn, both ice cream sales and crime rate increase at the same time (in summer months)
2. Use of causal connection
   – What inferences can be drawn from causal information and how?
   – e.g. what would change if the rooster were to cause the sun to rise, can we make the night shorter by waking him up early?

What is Special about these Questions?

- These are "What If?" kind of questions

- Interventional questions such as "What if I act?"
- Retrospective or explanatory questions such as "What if I had acted differently?"

- How would we answer such questions using the standard machine learning toolbox?

Discuss

Three Layer Causal Hierarchy

- We can think in terms of a classification of causal information
- Based on the type of questions that each class is capable of answering
- 3 – level hierarchy in the sense that questions at a level i (i = 1,2,3) can only be answered if information from a level j (j greater than or equal to i) is available

3-layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

[Pearl 2017]

3-layer Causal Hierarchy

Association: invokes purely statistical relationships, defined directly by the raw data

- This is learnt by any "black-box" of purely model free and data driven algorithm
- Famous examples such as that diapers and beer are often bought together

Intervention: ranks higher because it asks about a change in observed variables

- Example: what happens if we double the price – how will the customer respond?

3-level Causal Hierarchy

Counterfactuals: "What if I had acted differently?"

- Subsume interventional and associational questions

If we have a model at a higher level, the lower level can be answered easily

e.g., if we had counterfactual model, then the interventional question can be simply posed as:

> What would happen if we double the price? = What would happen *had the price been* double its current value?

Another Way to Conceptualize Hierarchy

- Action sentences
  - B (would be true) if we do A
- Counterfactuals
  - ¬B would change to B (B would be different) if it were A
- Explanation
  - B occurred because of A

Extended Version of Hierarchy

- Action sentences
  - B if we do A With probability p
- Counterfactuals
  - ¬B would change to B if it were A With probability p
- Explanation
  - B occurred because of A With probability p

Judea Pearl's Model: Major Ideas

Concept	Formalization
Causation	Encoding of behaviour under intervention
Intervention	Surgeries on mechanisms
Mechanisms	Functional Relationships by equations and graphs

Pearl's Model: Key Steps

- Devise a computational scheme for causality to facilitate prediction of the effects of "actions"
  - Use "Intervention" for "Action"
  - As actions are external entities originating "outside" the theory

- Mechanism: Autonomous physical laws or mechanisms of interest
  - We can change one without changing the others
  - e.g. logic gates of a circuit, mechanical linkages

Pearl's Model: Key Steps

- Intervention
  - Breakdown of a mechanism = surgery

- Causality
  - Which mechanism is to be surgically modified by a given action

Example to Ponder - 1

- If the grass is wet, then it rained
- If we break this bottle, the grass gets wet

- Conclusion: If we break this bottle, then it rained!

Example to Ponder - 2

- A suitcase will open *iff* both locks are open
- The right lock is open
- What happens if we open the left lock?

- Not sure – the right lock might get closed!

Modelling Causality

Causal Model M = (U, V, F)

- U = *Exo*genous variables
  - Values are determined by factors outside the model

- V = *Endo*genous variables
  - Values are described by structural equations

- F is a set of structural equations $\{F_X | X \in V\}$ (endogenous)
  - F_X is a mapping, tells us the value of X given the values of all the other variables in *U* and *V*
  - represents a mechanism or law in the world

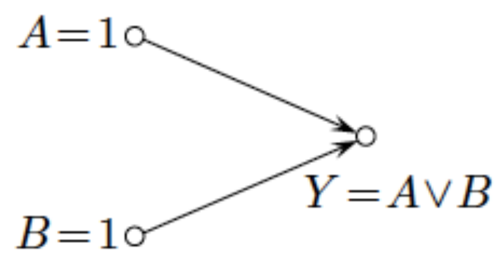
Example: Modelling Causality

- Forest fire could be caused by lightning or a lit match by an arsonist
- Endogenous variables, Boolean
  - F for fire
  - L for lightning
  - ML for match lit.
- Exogenous variables, U
  - Whether wood is dry
  - Whether there is enough oxygen in the air

$$F_F(U, L, ML) \text{ s.t. } F = 1 \text{ if } L = 1 \text{ or } ML = 1$$

Causal Networks

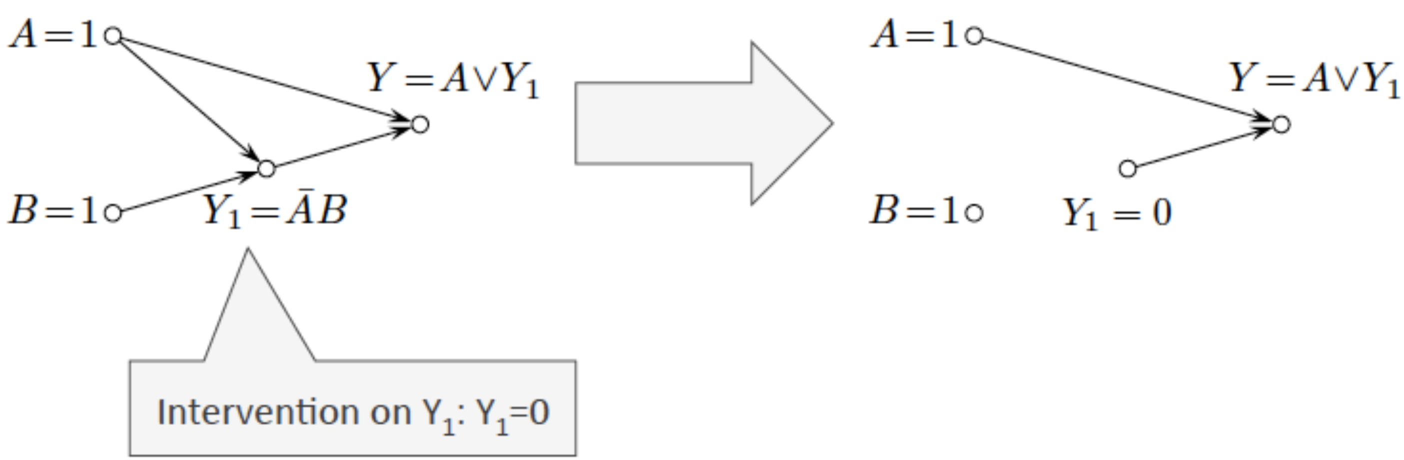
- Causal structural models:
  - Variables: A, B, Y
  - Structural equations: Y = A v B



- Modeling problems:
  - *E.g., A bottle breaks if either Alice or Bob throw a rock at it.*
  - Endogenous variables:
    - Alice throws a rock (A)
    - Bob throws a rock (B)
    - The bottle breaks (Y)
  - Exogenous variables:
    - Alice's aim, speed of the wind, bottle material etc.

Intervention/Contingency

- External interventions modify the structural equations or values of the variables.



Counterfactuals

- If ***not A*** then ***not φ***
  - In the absence of a cause, the effect doesn't occur

$$C = A \wedge B, \;\; A = 1 \wedge B = 1 \longleftarrow \text{Both counterfactual}$$

- Problem: Disjunctive causes
  - If Alice doesn't throw a rock, the bottle still breaks (because of Bob)
  - Neither Alice nor Bob are counterfactual causes

$$C = A \vee B, \;\; A = 1 \wedge B = 1 \longleftarrow \text{No counterfactual causes}$$

Actual Causes

[simplification]

A variable X is an <u>actual cause</u> of an effect Y if there exists a contingency that makes X counterfactual for Y.

$$C = A \vee B, \quad A = 1 \wedge B = 1 \longleftarrow$$ A is a cause under the contingency B=0

A Definition of Actual Cause

- Actual causes are of the form
  - $X_1 = x_1 \wedge X_2 = x_2 \wedge ..... \wedge X_k = X_k$
  - In short, X = x
- For X = x to be an actual cause of event Z The following three conditions should hold
  - Both X = x and Z are true in the actual world
  - Changing X to x' and some other variables W from w to w' changes Z from true to false
  - Setting W to w' does not have an effect on Z
- X is minimal- no subset of X satisfies the above two conditions

Example 1

$$Y = X_1 \wedge X_2$$

$X_1 = 1$ is counterfactual for $Y = 1$

$X_1 = 1, X_2 = 1 \Rightarrow Y = 1$

$X_1 = 0, X_2 = 1 \Rightarrow Y = 0$

Example 2

$$Y = X_1 \vee X_2$$

$X_1 = 1$ is not counterfactual for $Y=1$

$X_1 = 1$ is an actual cause for $Y = 1$, with contingency $X_2 = 0$

$X_1 = 1, X_2 = 1 \Rightarrow Y = 1$

$X_1 = 1, X_2 = 0 \Rightarrow Y = 1$

$X_1 = 0, X_2 = 0 \Rightarrow Y = 0$

Example 3

$$Y = (\neg X_1 \wedge X_2) \vee X_3$$

$X_1 = 1$ is not counterfactual for $Y = 1$

$X_1 = 1$ is not an actual cause for $Y = 1$

$X_1 = 1, X_2 = 1, X_3 = 1 \Rightarrow Y = 1$

$X_1 = 0, X_2 = 1, X_3 = 1 \Rightarrow Y = 1$

$X_1 = 1, X_2 = 0, X_3 = 1 \Rightarrow Y = 1$

$X_1 = 0, X_2 = 0, X_3 = 1 \Rightarrow Y = 1$

....

Y never changes by flipping X_1 for all combinations of X_2, X_3

Measure of Causality: Responsibility

A measure of the degree of causality

$$\rho = \frac{1}{1 + \min_{\Gamma} |\Gamma|}$$

size of the contingency set

Example

$Y = A \wedge (B \vee C)$ $\quad A = B = C = 1 \Rightarrow Y = 1$

A=1 is counterfactual for Y=1 (ρ=1)

B=1 is an actual cause for Y=1, with contingency C=0 (ρ=0.5)

Probabilistic Causal Model

Represented by a pair $(M, P(u))$

- $P(u)$ is a probability function defined over the exogenous variables U

- Each endogenous variable in V is a function of exogenous variables U
  - also gives a distribution on V

- In turn gives the probability of counter-factual statement $Pr(Y_{X=x} = y)$ or simply $Pr(Y_X = y)$

Probabilistic Model

Necessity

$$Pr(Y_{X=x'} = y' | X = x, Y = y) \\ = Pr(y'_{x'} | x, y)$$

The probability that event y would not have occurred in the absence of event x, (= $y'_{x'}$), given that x and y did in fact occur

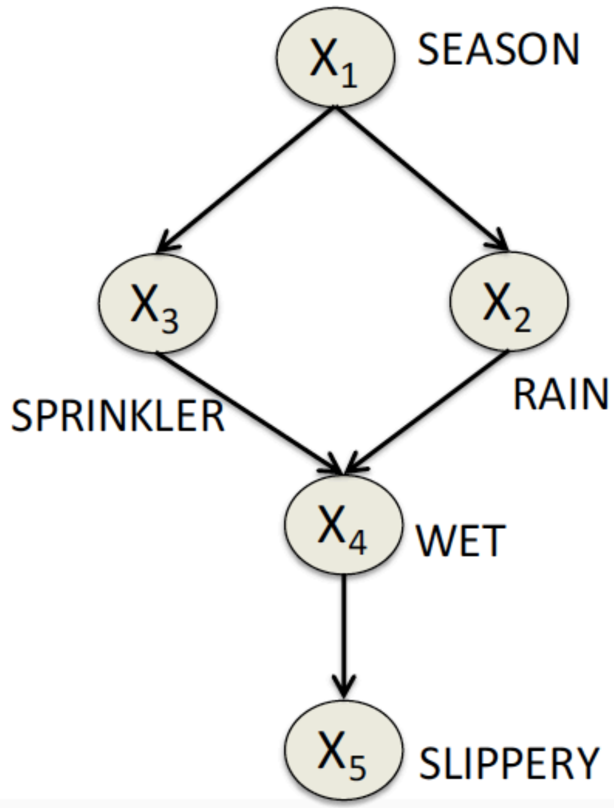
Sufficiency

$$Pr(Y_{X=x} = y | X = x', Y = y') \\ = Pr(y_x | x', y')$$

The probability that setting x would produce y in a situation where x & y are in fact absent

Ability of event x *to produce* event y

Worked Example on Structural Equations: Conditional Probability vs. Action



Observing versus Acting to make $X_3 = ON$

Conditional Probability of a Counterfactual Sentence

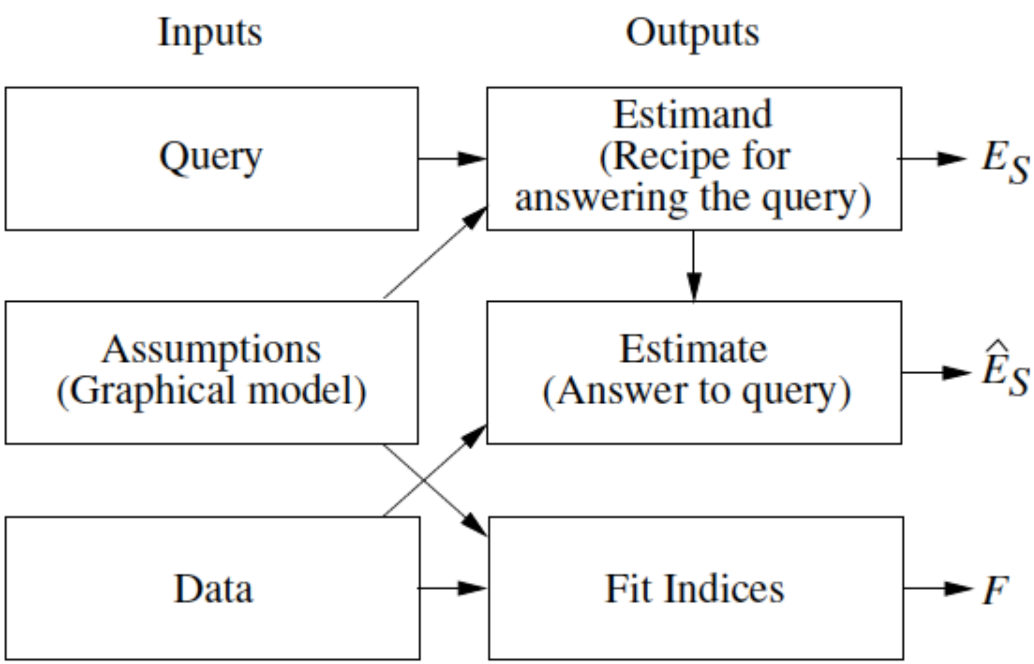
If we want to compute probability of:

“ {if it were A then B} given evidence e ”

we might use the following three step procedure:

1. Abduction
   – Update $P(u)$ by evidence to get $P(u|e)$
2. Action
   – Modify M by action $do(A)$, where A is antecedant of the counterfactual, to yield M_A
3. Deduction
   – Use $P(u|e)$ and M_A to compute probability of counterfactual consequence B

Pearl's View of a Structural Equations based "Inference Engine"



Answer to query

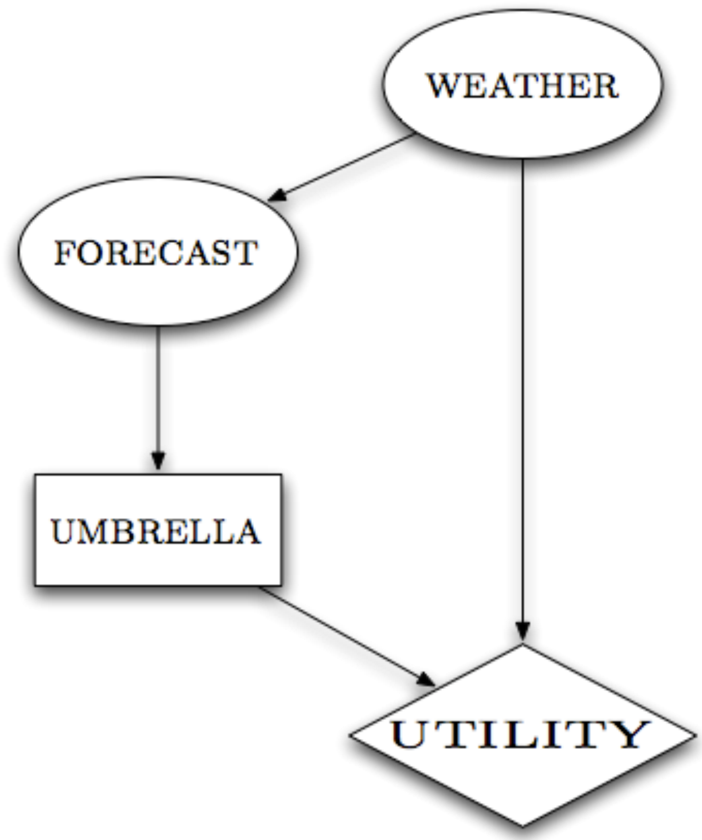
Answer +
estimated confidence

Fit of data to
model assumptions

[Pearl 2017]

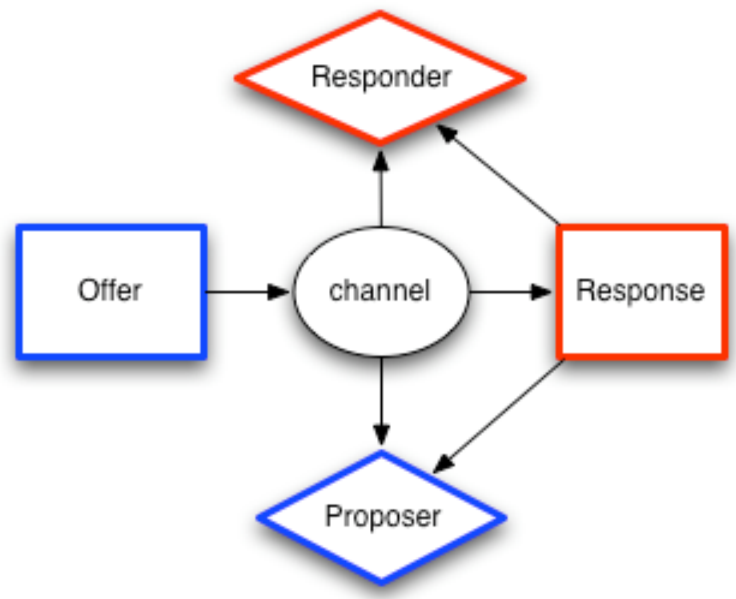
Recap: Influence Diagrams [Howard & Matheson '84]

- Influence Diagrams (ID) extend Bayesian Networks for decision making.
- *Rectangles* are decisions; *ovals* are chance variables; *diamonds* are utility functions.
- Graph topology describes decision problem.
- Each node specifies a probability distribution (CPD) given each value of parents.



Multi-agent Influence Diagrams
[Milch and Koller '01]

- Extend Influence Diagrams to the multi-agent case.
- Rectangles and diamonds represent decisions and utilities associated with agents; ovals represent chance variables.
- A strategy for a decision is a mapping from the informational parents of the decision to a value in its domain.
- A strategy profile includes strategies for all decisions.



Reasoning Patterns through IDs

- Informally, a reasoning pattern is a form of argument that leads to and explains a decision
  – e.g.
    - modus ponens in logic
    - explaining away in Bayes nets

- **What reasoning patterns can agents use in *interactive* decision making contexts?**

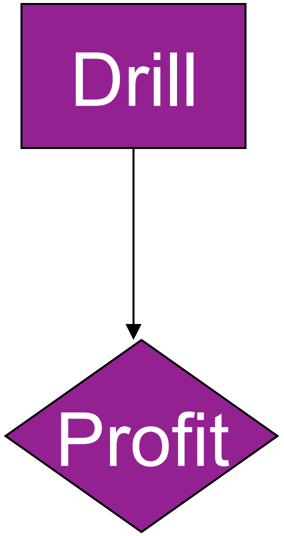
[A. Pfeffer & Y. Gal, On the reasoning patterns of agents in games, In Proc. AAAI 2007]

Characterization of Reasoning Patterns

- Four basic reasoning patterns, each characterized by paths in a multiple-agent version of influence diagrams

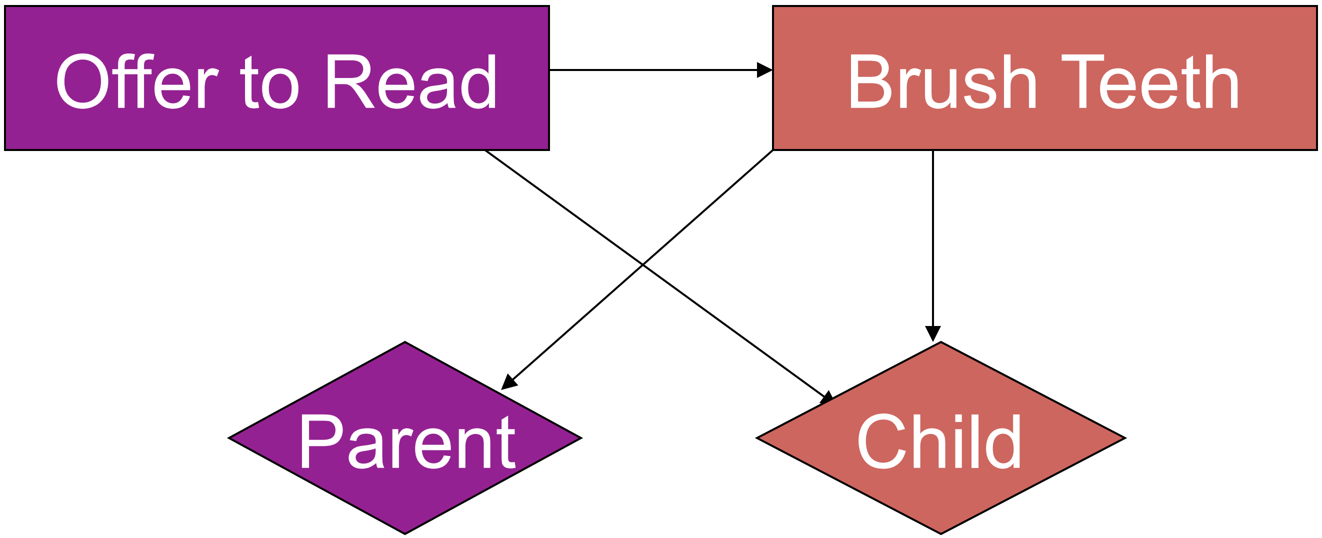
- Characterization based on graphical criteria only
  - could further refine characterization based on numerical parameters

Reasoning Pattern #1: Direct Effect



- An agent takes a decision because of its direct effect on its utility
  - without being mediated by other agents' actions

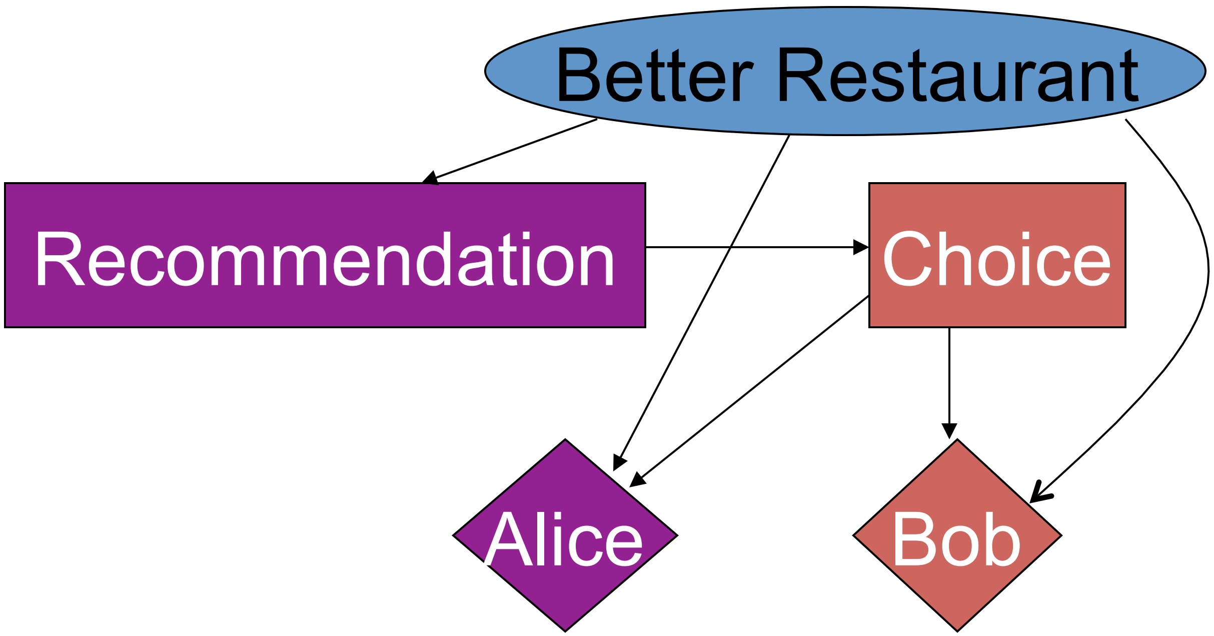
Reasoning Pattern #2: Manipulation



- Child knows about parent's action
- Parent does not care about reading, but wants child to brush teeth
- Child dislikes brushing teeth but likes being read to

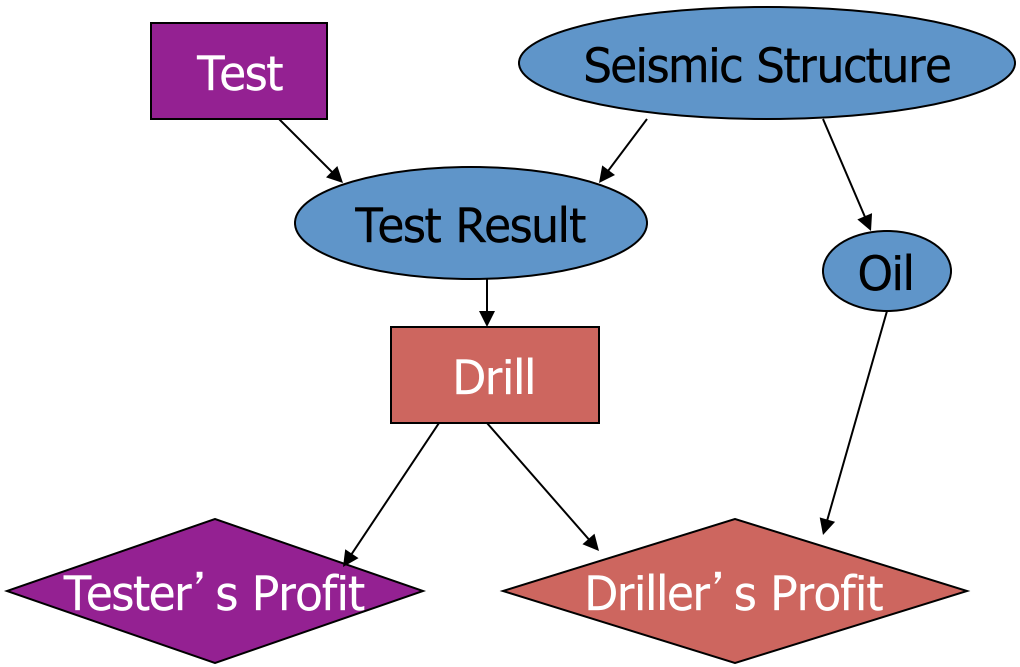
$\Rightarrow$ Parent can manipulate child

Reasoning Pattern #3: Signaling



- A communicates something that she knows to B, thus influencing B's behavior

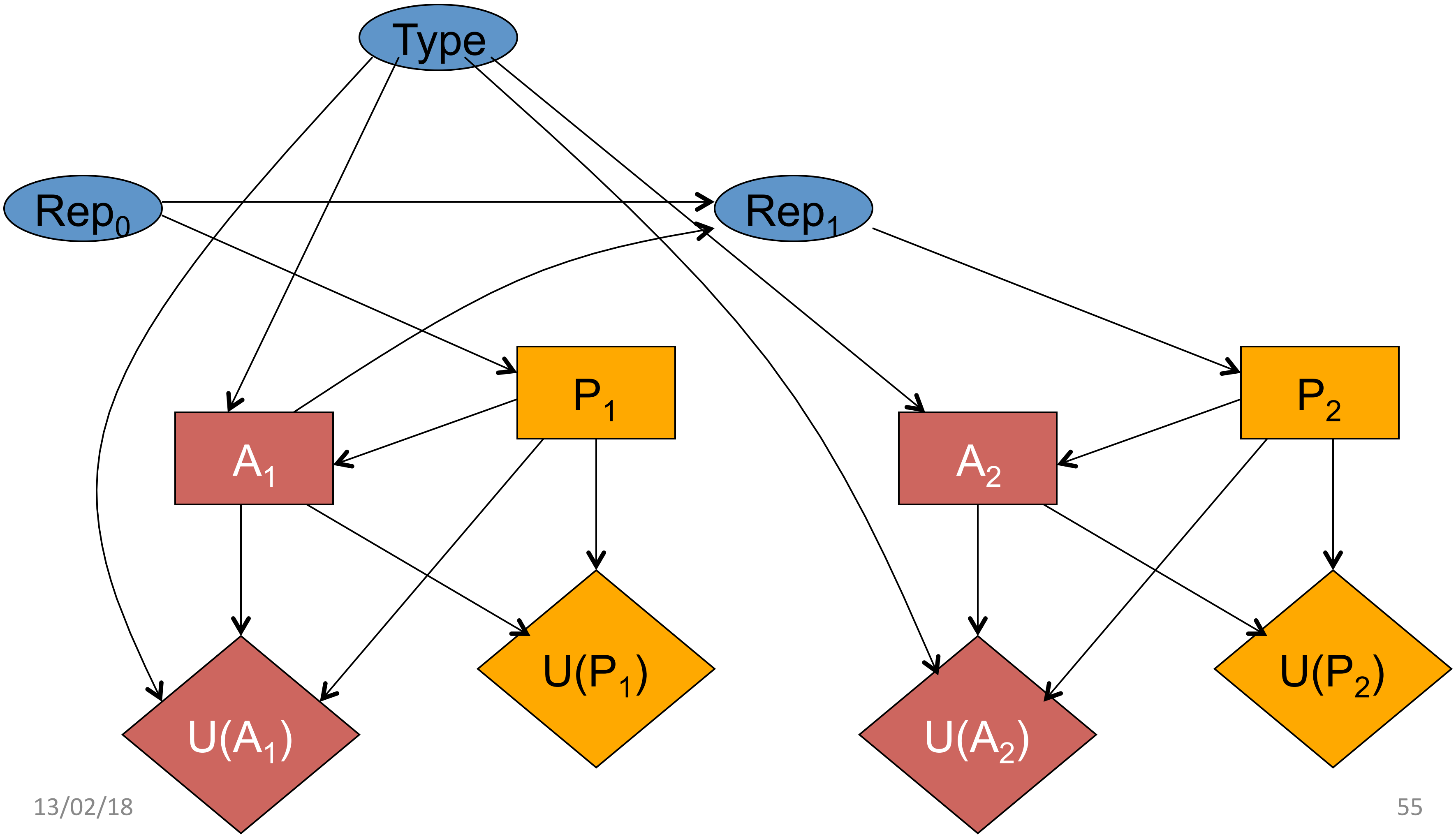
Reasoning Pattern #4: Revealing/Denying



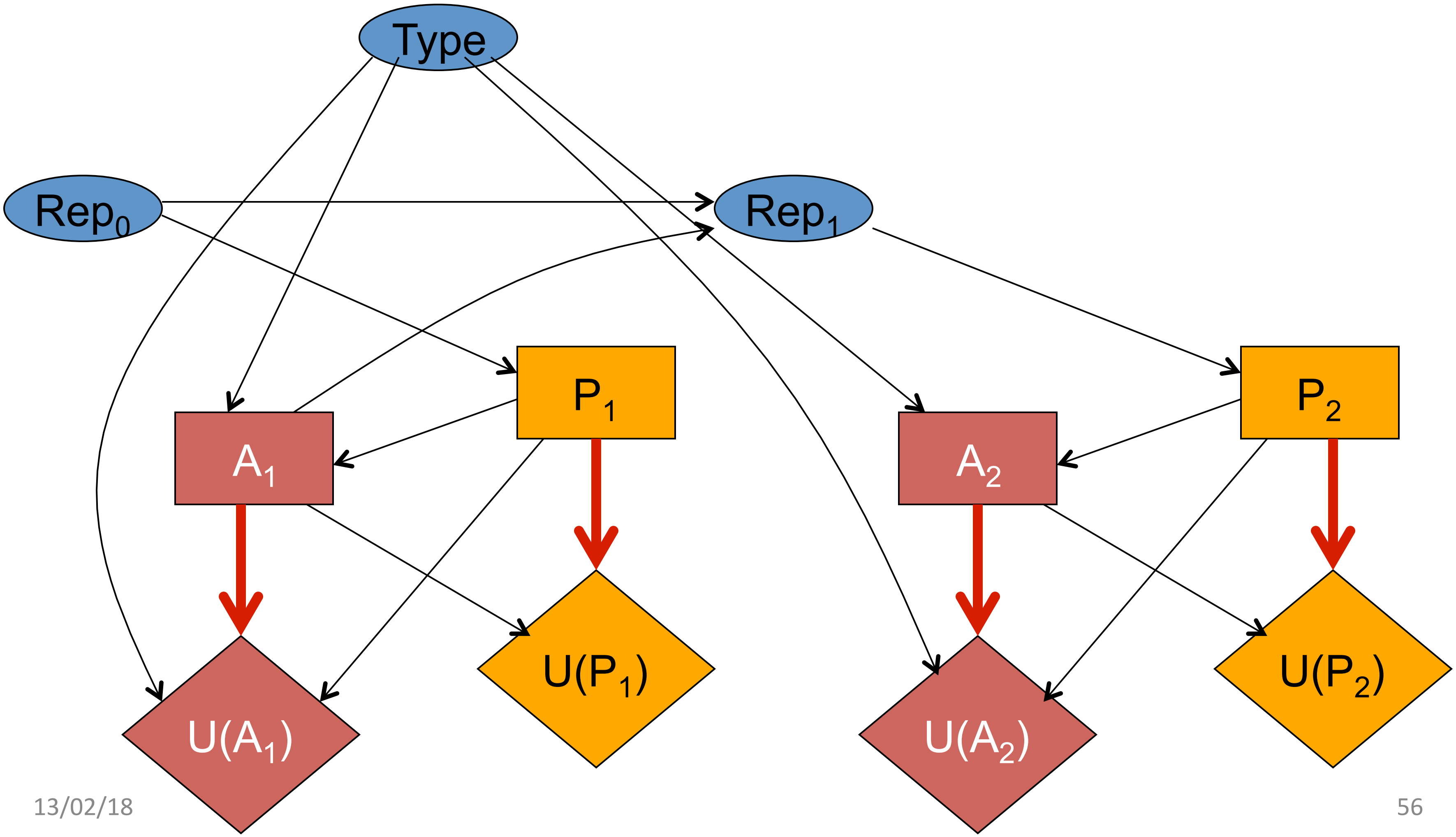
- Driller cares about oil
- Tester receives fee if driller drills
- Tester causes driller to find out (or not) about information tester herself does *not* know

Example: Two Stage Principal-Agent *Game*

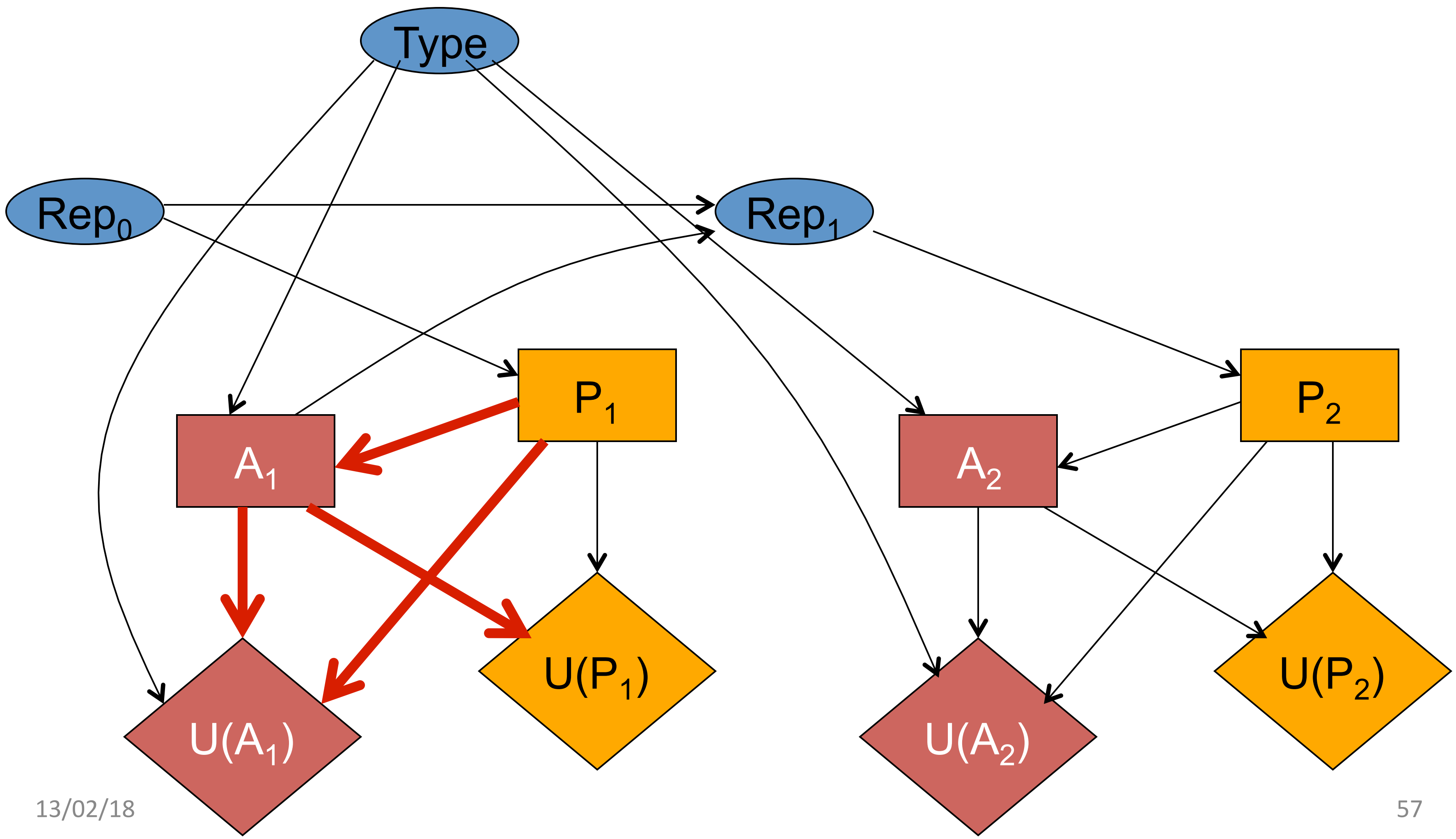
Type: described parameters specific to an agent
Rep: Quantification of “Reputation”



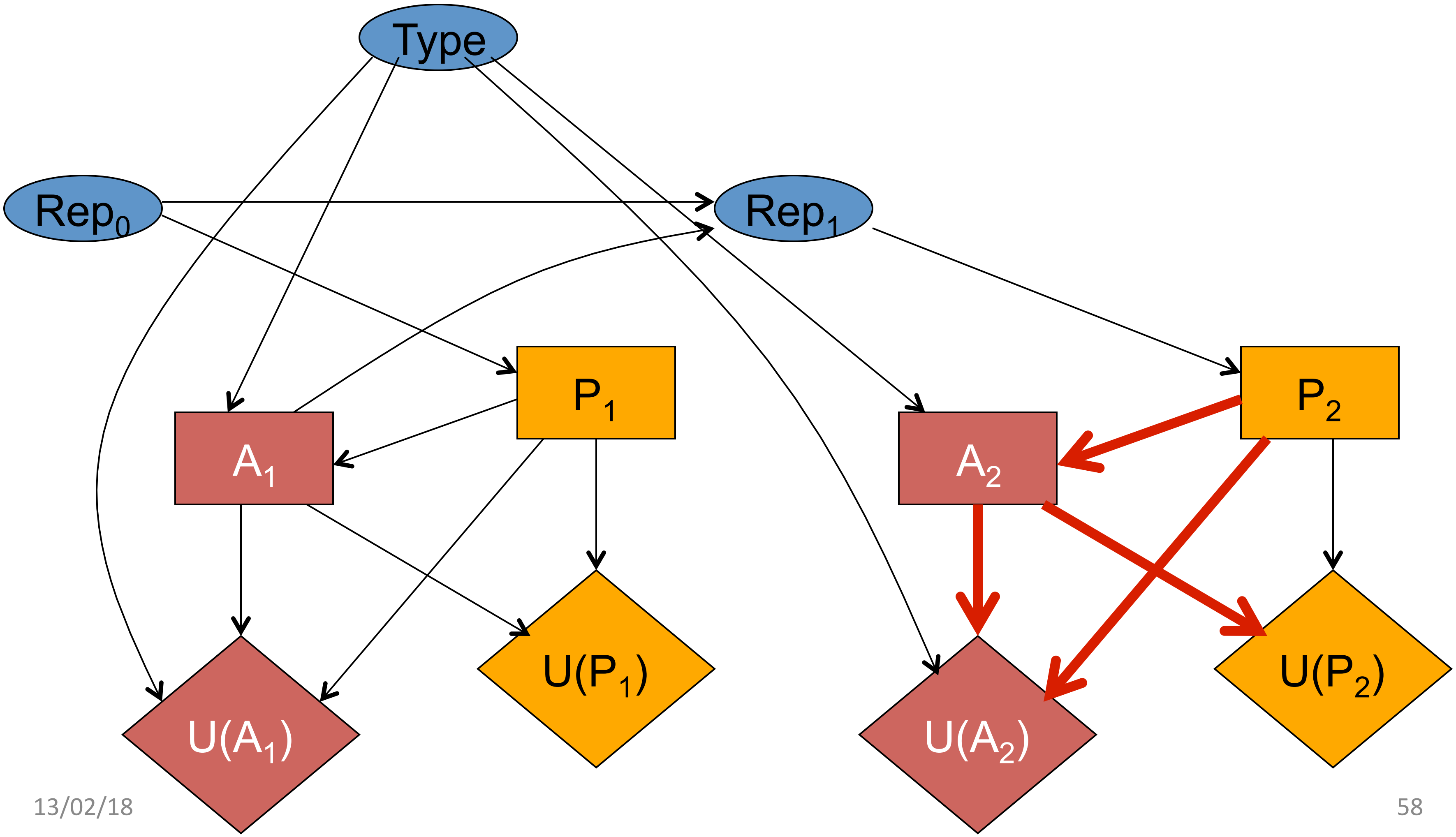
Direct Effect For All Four Decisions



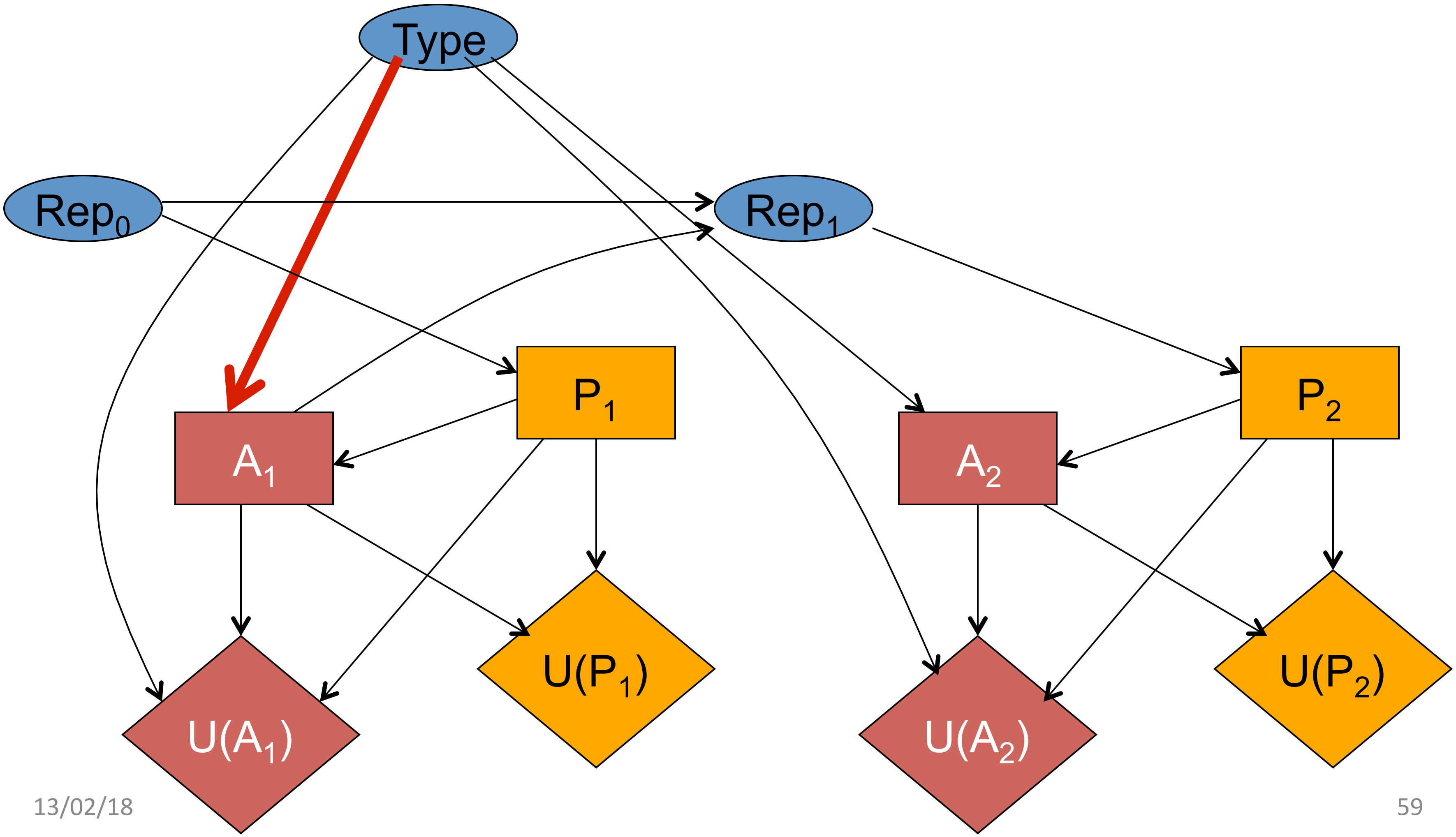
Manipulation ($P_1 \rightarrow A_1$)



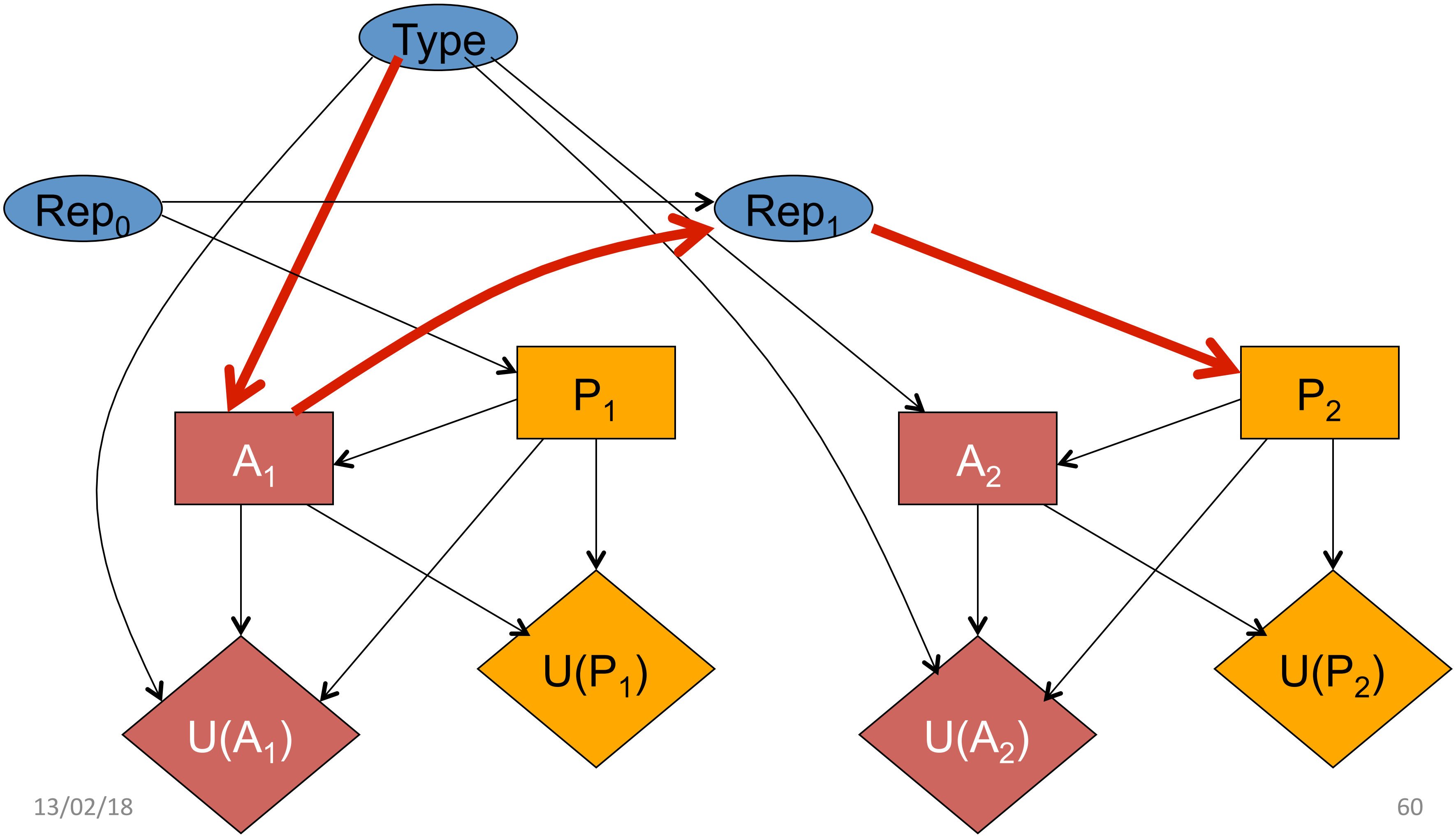
Manipulation ($P_2 \rightarrow A_2$)



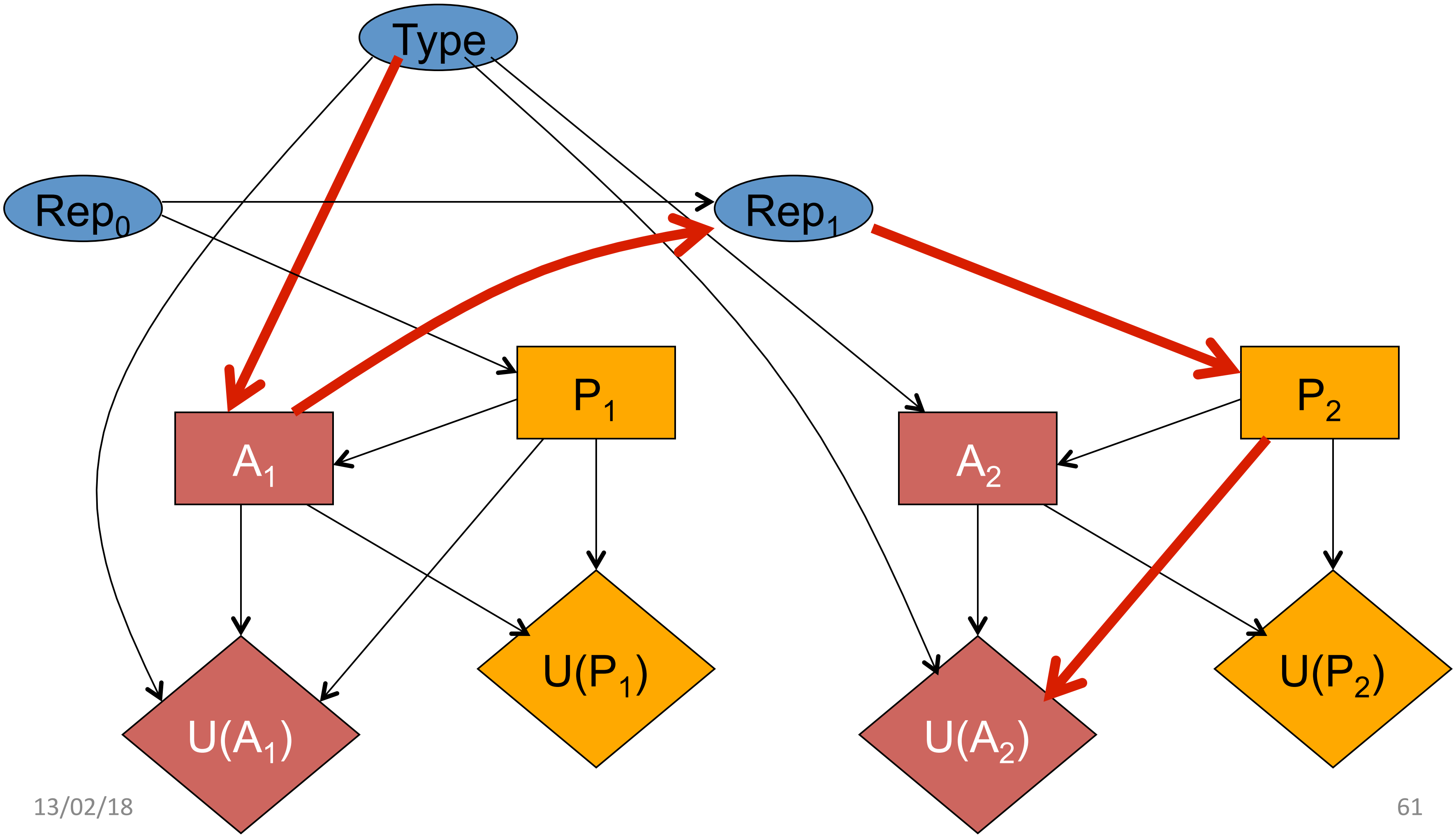
Signaling (A_1 signals Type to P_2)



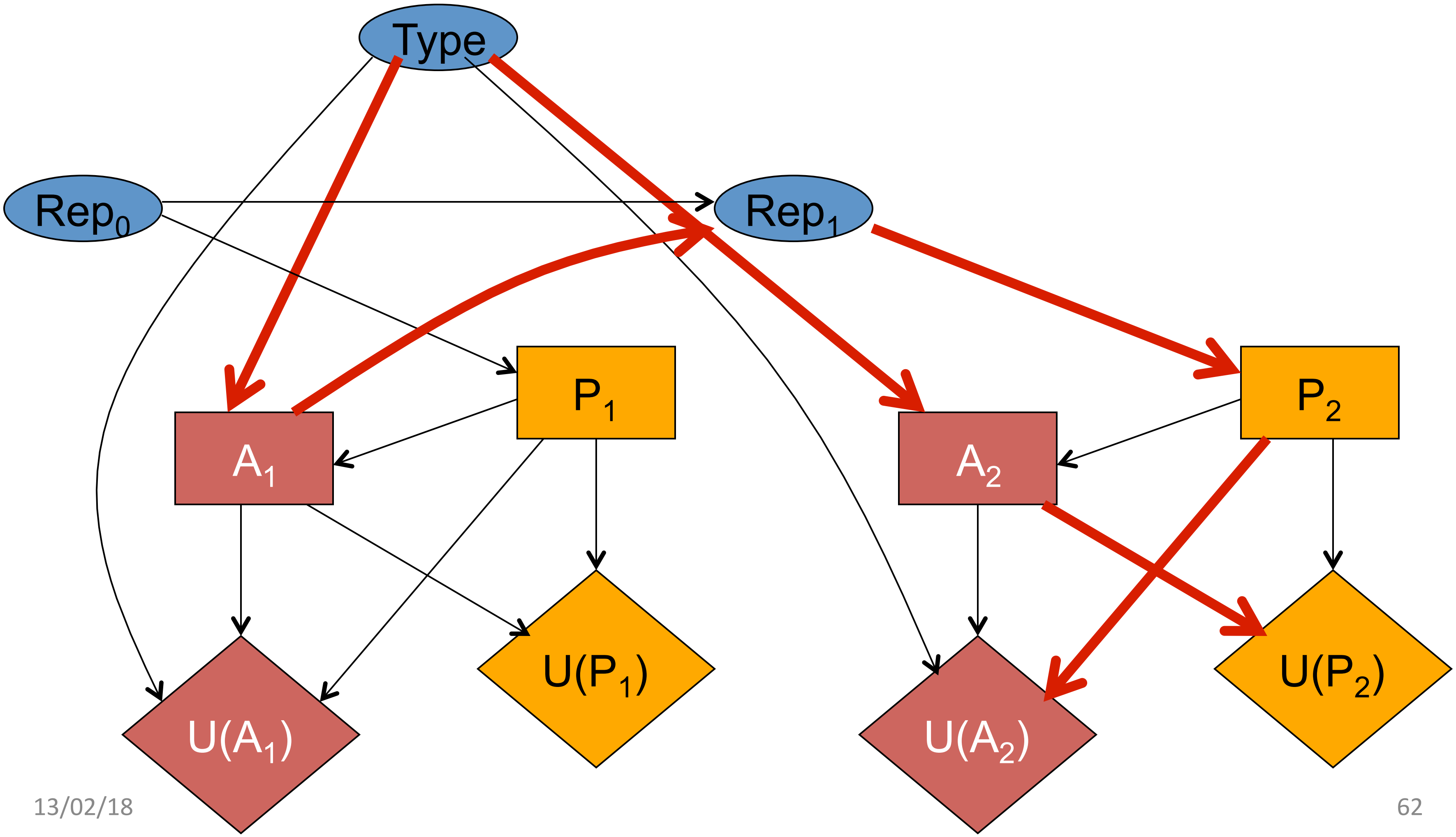
Signaling (A_1 signals Type to P_2)



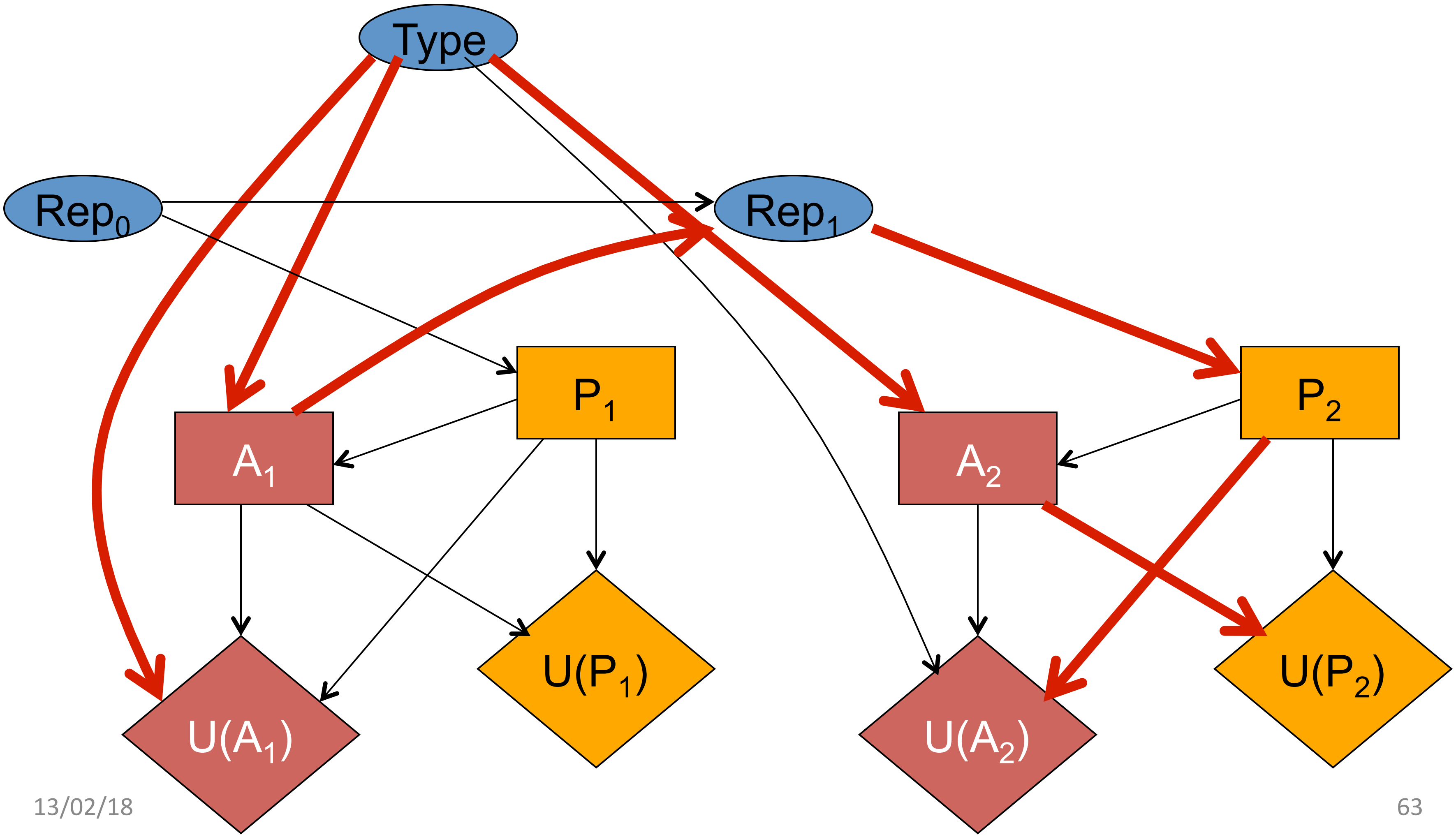
Signaling (A_1 signals Type to P_2)



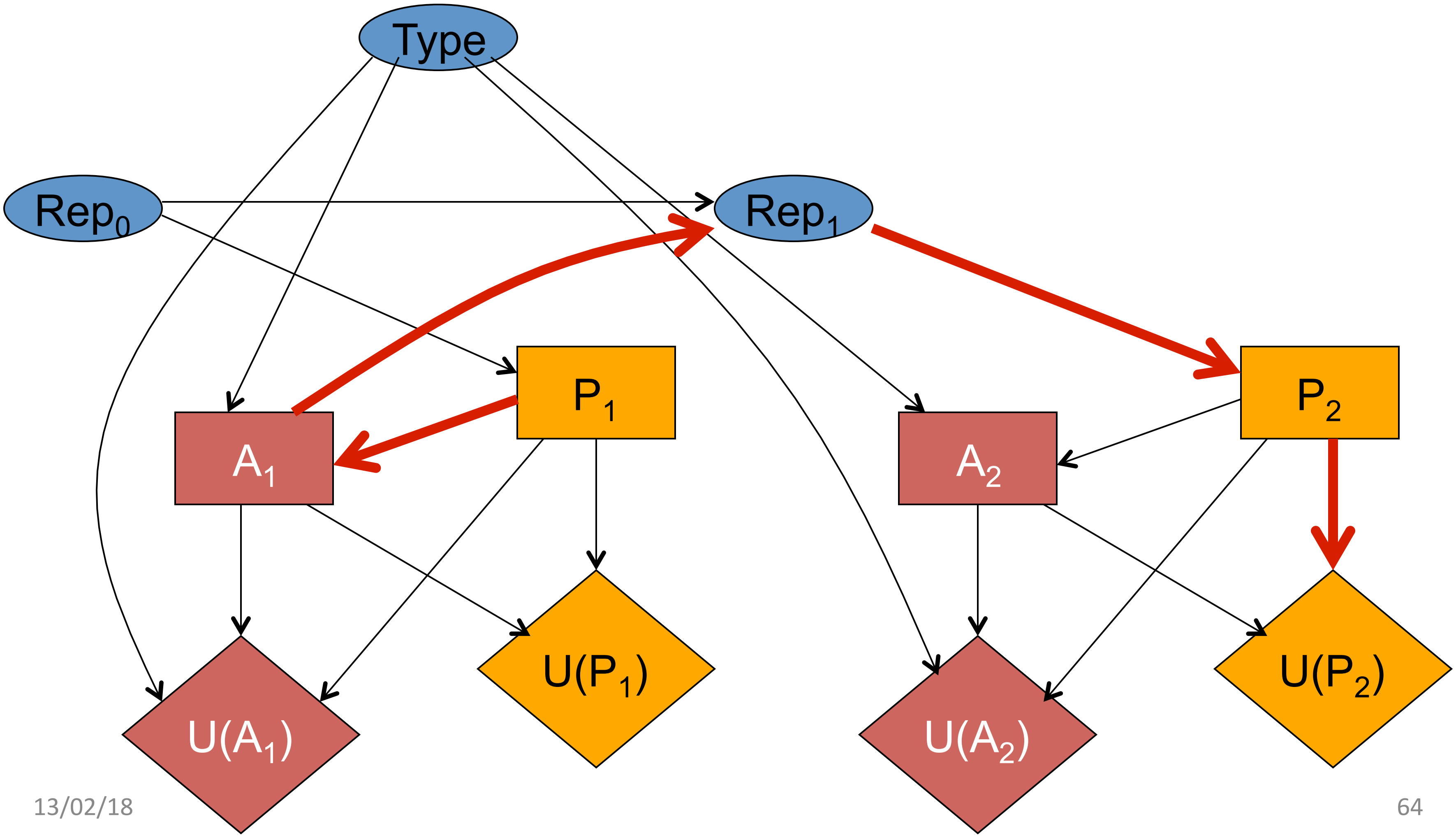
Signaling (A_1 signals Type to P_2)



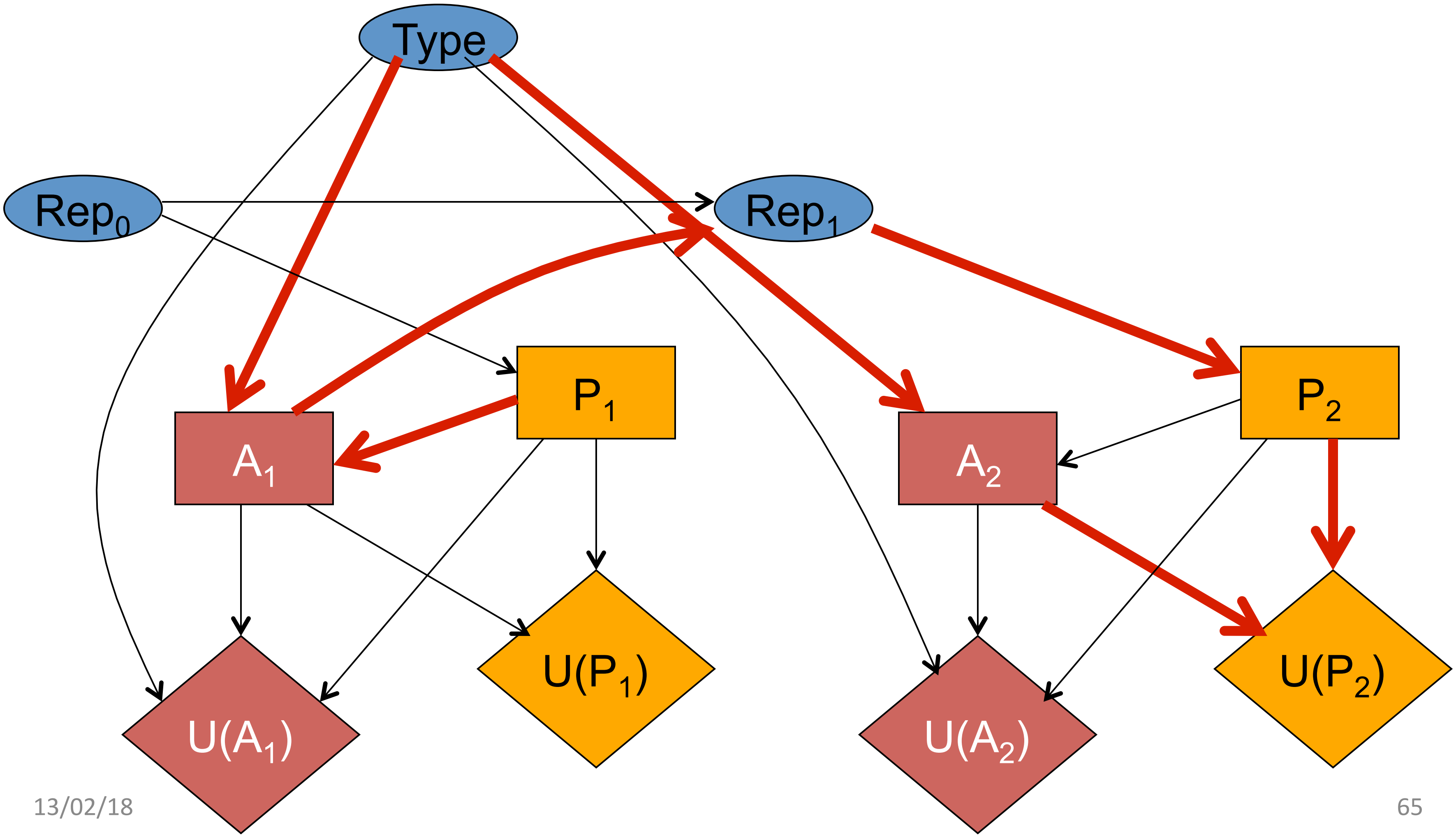
Signaling (A_1 signals Type to P_2)



Revealing/Denying (P_1 reveals Type to P_2)



Revealing/Denying (P_1 reveals Type to P_2)



Acknowledgement

The source of some of these slides is a VLDB 2014 tutorial entitled "Causality and Explanations in Databases", by Meliou, Roy, Suciu.